DOCUMENT RESUME

ED 025 319 PS 001 480

By- Alpern, Gerald D.; Levitt, Eugene E.

Methodoligical Considerations in Devising Head Start Program Evaluations.

Indiana Univ., Indianapolis. Medical Center.
Pub Date Apr 67
Note- 13p., Paper presented at the biennial meeting of the Society for Research in Child Development, April, 1967.
EDRS Price MF-$0.25 HC-$0.75
Descriptors- Control Groups, *Culturally Disadvantaged, Evaluation Methods, *Evaluation Techniques, Information Dissemination, Measurement Instruments, *Preschool Programs, *Program Evaluation, *Research Methodology, Testing Problems
Identifiers- *Head Start

In an attempt to improve Head Start evaluations, several methodological techniques are proposed. Since programs vary in approach, evaluations must be made on the success of the individual programs. Formulation of research questions should provide information as to the process and outcome of the program. To avoid experimenter bias, experimenters should be selected on the basis of their disengagement from Head Start. A baseline group (either a control group or the experimental group assessed on pretreatment performance should be used, and variables affecting their behavior should be noted. Some of the problems due to the lack of measuring instruments could be avoided if experimenters would not measure specific behavior as indicative of general ability. To avoid the problem of publishing only positive Head Start reports, the Office of Economic Opportunity should publish annually all Head Start evaluations. Several references are included. (JS)

U. S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED025319

Methodological Considerations in Devising Head Start Program Evaluations[1]

Gerald D. Alpern and Eugene E. Levitt

Indiana University Medical Center

This summer Head Start will celebrate its third birthday, thus leaving the toddler category and entering those formidable preschool years. Many people can remember the neonate; I (GDA) can clearly recall the pregnancy. In the spring of 1965, I was asked to help prepare novice preschool teachers for the first, eight-week, summer preschool program which represented the birth of Head Start in Indianapolis. I was allotted an hour and a half to provide some instruction in the psycho-social characteristics of the pre-school child. Minutes prior to my presentation, a collection of mimeographed sheets were thrust upon me along with the responsibility for tutoring the teachers-to-be in the procedures described on those pages. The contents of those sheets were a collection of evaluation procedures the majority of which suffered from gross methodological weaknesses. For example, the already overwhelmed teachers were going to be asked to provide measurements on their children using a number of spanking-new subjective scales whose inter-rater reliability could only be established with extensive training. There was no training at all for the teachers in the use of these measures. One of the other instruments was a well-standardized test with creditable reliability and validity. But you can imagine the usefulness of Goodenough data obtained from children in the midst of a newly established nursery school and scored by inexperienced people.

The idea that evaluation should be immediately built into a new program is praiseworthy in itself. But the value of that idea is largely

ED025319

PS001480

vitiated when there are serious methodological deficiencies in the evaluation plan. A survey of many of the published and mimeographed reports of Head Start evaluations suggests that many of them suffered from such methodological shortcomings. The purpose of this paper is to elucidate some major methodological considerations, to identify potential errors with the hope that future Head Start evaluations may be improved. It is our view that so-called field conditions rarely provide a legitimate excuse for poor research methodology. As we shall point out shortly, Head Start programs are characterized by considerable heterogeneity so that our comments will not be relevant to every program. We hope that the relevancy will extend at least to a majority.

Our presentation is subsumed under five headings:

1. Formulation of Research Questions
2. The Problem of Investigator Bias
3. The Problem of Sampling
4. The Problem of Appropriate Measuring Instruments
5. The Problem of Data Dissemination

Formulation of Research Questions

Approximately 560,000 children at 13,400 centers participated in the original Head Start projects in the summer of 1965. The heterogeneity of these programs and their evaluation approaches is probably only passingly reflected in a variety of available professional and popular reports. Hechinger (1965) quotes Martin Deutsch as noting that the original Head Start programs ranged from "excellent to purely custodial." Diversity has greatly increased since 1965. The original eight-week nursery school

scheme has been expanded to include year-round centers, medical, dental and dietary programs, and parent involvement projects ranging from occasional discussion meetings to on-going community action groups. The original idea of preschool education has been thoroughly modified in some centers to the point where major emphasis is now on parent education, rather than directly on the child. Even those centers which continue with a primary focus on a type of nursery school "enrichment" environment run the gamut of preschool philosophies. The point is that Head Start is clearly not a single program. No one study could possibly demonstrate its value.[2] All that can reasonably be asked is to estimate the effect that a particular program has on a particular population with reference to a particular variable or variables. For example, what are the effects of a six-month, Montesorri-type program on the reading readiness of 4-year-old underprivileged children at the time of entrance to first grade? Or, what are the effects of a year of weekly parent discussions on the achievement motivation of their first-grade children? These questions, of course, require definitional amplification. They illustrate the kind of initial specificity which is a prerequisite to experimental formulation.

These illustrative questions concern outcome. Because the Head Start program is not unitary, every project must be individually evaluated from the point of view of outcome. However, to restrict research attention solely to outcome would lose a potentially valuable body of information. We refer here to the data which would be derived from questions concerning process rather than outcome. They are the "why" questions, as distinct from the "what" questions of outcome. For example, are patterns of teacher behavior relevant to outcome, as suggested in the study of Conners and Eisenberg

(1966) in Baltimore?

Some of these questions can be answered by appropriate correlational analyses of intraproject evaluation data. Some can be answered by a comparison of outcome data among different Head Start programs.

Many current Head Start programs have multifaceted approaches, which involve various attempts to influence both children and parents. The conventional outcome question concerns the aggregate effect of all these influences on child performance. This is primarily an applied question; the applied social scientist is concerned mostly to discover if a particular treatment is effective. He is not always concerned with what makes it effective. But it may be of considerable value, both pure and applied, to isolate the effects of individual program aspects. Ideally, this is accomplished by systematically varying the program so that only one of the facets operates in a given period of time. Or, the variation may take place among different centers within the same program during the same time period. Politically, neither of these approaches is likely to be feasible. That is, the ultimate purpose of Head Start is to alter behavior, not to act as a laboratory within which information may be obtained for pure science purposes. Depriving a Head Start operation or parts of the operation of some beneficial influences for purely investigative reasons is likely to encounter severe community resistance.

An alternative is to obtain the necessary information by comparing local Head Start programs with different curricula and approaches. There are obvious methodological problems; local programs will differ among themselves along dimensions other than curricula. However, at least

tentative conclusions would be available if there were enough local programs so that interprogram differences other than in curriculum would be counterbalanced and thus neutralized.

The Problem of Investigator Bias

The basic reference on the effect of the experimenter on his own results is Robert Rosenthal's classic volume, Experimenter Effects in Behavioral Research (1966). We can no longer question the contention that any investigator, no matter how well grounded he may be in scientific methodology, may unwittingly exercise an extraneous influence on his findings, an influence which obfuscates the data and leads to erroneous conclusions. Surely there are few research areas which can compare with Head Start in potential for experimenter bias. Head Start has powerful implications for a social philosophy as well as a political stance. Many of the individuals involved in Head Start are true partisans, believers, individuals whose involvement reflects their dedication to a philosophy. Furthermore, we must consider that one of the practicalities of the political arena is that funding of a Head Start program--the continuation of its very existence--may depend upon a particular evaluation.

We do not claim that investigator bias has as yet been identified in any Head Start evaluation. However, some gross methodological shortcomings suggest the operation of bias, influencing investigators who ought to know better. Going no further than a review of literature sections, a sizable proportion of Head Start research reports indicate a conspicuous absence of reference to those studies which have furnished negative findings. Statistical errors are common as, for example, in a report of

PS001480

a Head Start program in the Los Angeles area (Garwood and Augenbraum, 1966). Many statistical comparisons were presented in a table which carried the following footnote: "All results in the predicted direction were tested by one-tailed tests. Two-tailed tests were used for those cases where results were in the opposite direction to that predicted." Was the experimenter unaware that when a one-tailed test is contemplated, a difference in the direction opposite to that predicted is identical with zero, and cannot be subjected to any statistical test?

If we accept that Head Start provides a fertile field for experimenter bias, it is of the greatest importance to control this bias to the maximum feasible degree. Probably the most effective single method is by experimenter selection. It follows that not everyone should be permitted to evaluate Head Start programs, certainly not individuals directly associated with local programs. Minimization of investigator bias requires that the evaluator be emotionally as well as actually disengaged from Head Start, that he should have no preconceptions or attitudes which might affect experimental findings, and no personal stake in the outcome of evaluation.[3] Rosenthal's concept of the "professional experimenter" is pertinent.

The Problem of Sampling

We need not belabor the point that any demonstration of the effectiveness of a program requires a baseline whose purpose is to estimate the extent of change which can be attributed to factors other than the program itself. The baseline is a requisite whether we deal with therapy, teaching, or attitude change. It is most necessary when the subject is a young child, still relatively plastic and developing, for whom both intrinsic and

extrinsic factors are more influential in effecting change than for the adult. A baseline can be provided either by a control group, or by an assessment of pre-treatment performance in the experimental group. When the latter method is used, subjects may be lost over time due to family mobility. Some accounting of this subject attrition is absolutely necessary since those who are most mobile may reflect a homogenous segment of the experimental group, such as those who score lower on most evaluation instruments. The sample which remains at the end of a longitudinal study may not be typical of the original sample. For this reason, it is necessary to compare the samples at the outset and at the conclusion of a longitudinal study on as many background variables as are available. If a number of differences are demonstrated, great caution must be exercised in drawing conclusions from the data.

The longitudinal investigation with its own-control must be employed judiciously. Primarily, it is necessary to be able to assume with great safety that the dependent variable measures are not subject to change over time in the absence of specific influences of the Head Start program. This assumption is rarely tenable when we measure <u>general</u> abilities in young children. It is somewhat more likely to be tenable when a <u>specific</u> skill is being evaluated. In other instances, a control group matched for age, socio-economic status, race, sex, intelligence, and initial scores on dependent variables, is much superior to the own-control technique. Large Head Start programs may have waiting lists from which appropriate control subjects may be drawn and matched with children in the program.

Another possibility is the exploitation of chance differences in

independent variables within a program. To illustrate, suppose a particular program has four classes. It is quite possible that some kind of independent variable differences will accidentally exist among those classes, such as variable parent participation, differences in teacher attitude or experience, and so forth. Reliable differences in the dependent variables among the classes could be reasonably attributed to these chance differences.

The employment of an appropriate control group does not necessarily settle the problem of the baseline completely. There are certain other effects which need to be considered, though it may not always be possible to actually control for them. One is the so-called Hawthorne effect. Simply being in a treated group, rather than the influence of a specific program, may yield behavioral change. This can be controlled by providing irrelevant training for the control group. A second confounding effect comes about when subjects in control groups are drawn from the same neighborhood as experimental groups. In interacting under ordinary circumstances in the neighborhood, members of the control group may be favorably affected through contact with members of the experimental group who have been exposed to the program. This so-called diffusion effect could obfuscate real experimental-control differences which are due directly to the program. It can be controlled by the use of a "distal control group" as in the study of Gray and Klaus (1965).

Practical considerations in the field may prevent all of these factors from being included in the experimental design, but it is well for the experimenter to be aware that they may influence his results and that he

should control for them whenever it is possible.

The Problem of Measuring Instruments

One of the complaints of Head Start evaluators is that there is a paucity of available measuring instruments which are appropriate for the preschool child. Consequently, new or ad hoc tools whose reliability and validity are unknown have been employed. Conclusions concerning programs have been based on findings with such unproven instruments.

The psychological significance of an increase in scores on an unproven measuring instrument is always debatable. There is nothing wrong with using a Head Start program to develop new measuring instruments, especially when public school adjustment or performance is used as the ultimate criterion against which to validate the new instrument. However, this is clearly different from using the ad hoc measure to evaluate a Head Start program.

The argument that established instruments are not available is itself open to question. In fact, there is a multitude of standardized tests for preschool children measuring all phases of cognitive processes (e.g. Wechsler Preschool and Primary Scale of Intelligence), motor and perceptual skills (e.g. Kephart's Motor Survey Test), language development (e.g. Templin Language Tests), preschool academic achievement (e.g. Metropolitan Readiness Test), and, of course, actual public school performance itself.

In most instances, the evaluators wish to assess a general ability rather than a specific behavior. Behavior on a particular test is used as an operational definition of the construct which carries the general

ability label. For example, we might use the ability to manipulate Wechsler blocks as a measure of a construct which we call "manual dexterity," though we recognize that there are other performances which might also be subsumed under this construct heading. A common mistake is to select as a construct definition, a behavior for which specific training has been given in the preschool program. In such an instance, it would not be surprising that the experimental group shows superior performance as compared to a control group. But it is scientifically incorrect to assume that generalization from a specifically practiced ability has extended to other abilities which logically fall in the same class. Again, there is nothing intrinsically unscientific in evaluating the outcome of a specific training. The error lies in using that specific behavior as a measure of a construct, in drawing conclusions about the effect of specific training on a general ability.

The Problem of Data Dissemination

We have expressed our belief that every Head Start program needs to be scientifically evaluated. We also believe that the results of every evaluation need to be presented fully so as to be available to all persons involved in Project Head Start. It is our view that it is unwise to leave this vital dissemination of information to the whim of professional journal editors. There is simply too much emphasis on the positive result in today's overcrowded journals. The investigation with negative findings, no matter how methodologically sound it may be, appears to have considerably less chance of appearing in conventional print. There is, furthermore, a reluctance on the part of some investigators to submit negative studies for publication.

In addition, the journal editor is as likely as not to be victimized by a fallacy which we pointed out earlier, namely, that it is possible to evaluate the Head Start program. He may believe that previous publication of Head Start evaluations is a basis for rejecting the manuscript at hand.

We propose as a solution the establishment of a collected volume of Head Start evaluations to be published annually by the Office of Economic Opportunity . It should be required that every Head Start evaluation be submitted for publication as soon as it is in report form. We see no alternative to such a volume if necessary dissemination of Head Start evaluations is to be accomplished.

In summary, we wish to point out again that the present paper is not a blanket condemnation of the Head Start program evaluations which have already been accomplished. Rather, we view it as having emanated from a sorting out of the more fruitful efforts from the less fruitful ones, in a search for underlying methodological procedures that distinguish the former group from the latter. It is our hope that this paper may contribute in some small way to increasing the efficiency with which Head Start programs are evaluated.

Footnotes

1 A paper presented in the symposium "Considerations in Evaluating Head Start Programs" at the biennial meeting of the Society for Research in Child Development, April, 1967.

2 Of course, if a large majority of Head Start programs around the country were determined to be effective, then we could reason inductively that Head Start is successful. This position is already established if we refer only to a specific interest of many Head Start programs--the enhancement of functional intelligence as measured by a standard test. There is no doubt that test intelligence is educable. This fact has been established for years, and is certainly not unique to Head Start programs. The interested reader is referred to Hunt's (1964) review.

3 Of course, it is possible that we are talking about a hypothetical individual who does not actually exist. Perhaps our point here is really that Head Start evaluators ought to be evaluated before they evaluate Head Start.

References

Conners, C. K. and Eisenberg, L. The effect of teacher behavior on verbal intelligence in Operation Head Start children. Unpublished manuscript, Johns Hopkins University School of Medicine, 1966.

Garwood, D. S. and Augenbraun, B. The Head Start child: relationship between psychosocial maturity and familial factors. Unpublished manuscript, Center for Early Education, Los Angeles, 1966.

Gray, S. W. and Klaus, R. A. An experimental preschool program for culturally deprived children. Child Development, 1965, 36, 887-898.

Hechinger, F. M. Head Start to where? The Saturday Review, 1965, 48, 58-60, 75.

Hunt, J. McV. The psychological basis for using pre-school enrichment as an antidote for cultural deprivation. Merrill-Palmer Quarterly, 1964, 10, 209-248.

Rosenthal, R. Experimenter Effects in Behavioral Research. New York: Appleton-Century-Crofts, 1966.